Phrase Table Induction Using In-Domain Monolingual Data for Domain Adaptation in Statistical Machine Translation

Benjamin Marie **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{bmarie, atsushi.fujita}@nict.go.jp

Abstract

We present a new framework to induce an in-domain phrase table from in-domain monolingual data that can be used to adapt a general-domain statistical machine translation system to the targeted domain. Our method first compiles sets of phrases in source and target languages separately and generates candidate phrase pairs by taking the Cartesian product of the two phrase sets. It then computes inexpensive features for each candidate phrase pair and filters them using a supervised classifier in order to induce an in-domain phrase table. We experimented on the language pair English–French, both translation directions, in two domains and obtained consistently better results than a strong baseline system that uses an in-domain bilingual lexicon. We also conducted an error analysis that showed the induced phrase tables proposed useful translations, especially for words and phrases unseen in the parallel data used to train the general-domain baseline system.

1 Introduction

In phrase-based statistical machine translation (SMT), translation models are estimated over a large amount of parallel data. In general, using more data leads to a better translation model. When no specific domain is targeted, *general-domain*[1] parallel data from various domains may be used to train a general-purpose SMT system. However, it is well-known that, in training a system to translate texts from a specific domain, using *in-domain* parallel data can lead to a significantly better translation quality (Carpuat et al., 2012). Indeed, when only general-domain parallel data are used, it is unlikely that the translation model can learn expressions and their translations specific to the targeted domain. Such expressions will then remain untranslated in the in-domain texts to translate.

So far, in-domain parallel data have been harnessed to cover domain-specific expressions and their translations in the translation model. However, even if we can assume the availability of a large quantity of general-domain parallel data, at least for resource-rich language pairs, finding in-domain parallel data specific to a particular domain remains challenging. In-domain parallel data may not exist for the targeted language pairs or may not be available at hand to train a good translation model.

In order to circumvent the lack of in-domain parallel data, this paper presents a new method to adapt an existing SMT system to a specific domain by inducing an in-domain phrase table, i.e., a set of phrase pairs associated with features for decoding, from in-domain monolingual data. As we review in Section 2, most of the existing methods for inducing phrase tables are not designed, and may not perform as expected, to induce a phrase table for a specific domain for which only limited resources are available. Instead of relying on large quantity of parallel data or highly comparable corpora, our method induces an in-domain phrase table from unaligned in-domain monolingual data through a three-step pro-

[1] As in Axelrod et al. (2011), in this paper, we use the term *general-domain* instead of the commonly used *out-of-domain* because we assume that the parallel data may contain some in-domain sentence pairs.

Transactions of the Association for Computational Linguistics, vol. 5, pp. 487–500, 2017. Action Editor: Chris Quirk.
Submission batch: 3/2017; Published 11/2017.

cedure: phrase collection, phrase pair scoring, and phrase pair filtering. Incorporating our induced in-domain phrase table into an SMT system achieves substantial improvements in translating in-domain texts over a strong baseline system, which uses an in-domain bilingual lexicon.

To achieve this improvement, our proposed method for inducing an in-domain phrase table addresses several limitations of previous work by:

- dealing with source and target phrases of arbitrary length collected from in-domain monolingual data,
- proposing translations for not only unseen source phrases, but also those already seen in the general-domain parallel data, and
- making use of potentially many features computed from the monolingual data, as well as from the parallel data, in order to score and filter the candidate phrase pairs.

In the remainder of this paper, we first review previous work in Section 2, highlighting the main weaknesses of existing methods for inducing a phrase table for domain adaptation, and our motivation. In Section 3, we then present our phrase table induction method with all the necessary steps: phrase collection (Section 3.1), computing features of each phrase pair (Section 3.2), and pruning the induced phrase tables to keep their size manageable (Section 3.3). In Section 4, we describe our experiments to evaluate the impact of the induced phrase tables in translating in-domain texts. Following the description of the data (Section 4.1), we explain the tools and parameters used to induce the phrase tables (Section 4.2), our SMT systems (Section 4.3), and present additional baseline systems (Section 4.4). Our experimental results are given in Section 4.5. Section 5.1 analyzes the error distribution of the translations produced by an SMT system using our induced phrase table, followed by translation examples to further illustrate its impact in Section 5.2. Finally, Section 6 concludes this work and proposes some possible improvements to our approach.

2 Motivation

In machine translation (MT), words and phrases that do not appear in the training parallel data, i.e., *out-of-vocabulary* (OOV) tokens, have been recognized as one of the fundamental issues, regardless of the scenario, such as adapting existing SMT systems to a new specific domain.

One straightforward way to find translations of OOV words and phrases consists in enlarging the parallel data used to train the translation model. This can be done by retrieving parallel sentences from comparable corpora. However, these methods heavily rely on document-level information (Zhao and Vogel, 2002; Utiyama and Isahara, 2003; Fung and Cheung, 2004; Munteanu and Marcu, 2005) to reduce their search space by scoring only sentence pairs extracted from each pair of documents. Indeed, scoring all possible sentence pairs from two large monolingual corpora using costly features and a classifier, as proposed by Munteanu and Marcu (2005) for instance, is computationally too expensive.[2] In many cases, we may not have access to document-level information in the given monolingual data for the targeted domain. Furthermore, even without considering computational cost, it is unlikely that a large number of parallel sentences can be retrieved from non-comparable monolingual corpora. Hewavitharana and Vogel (2016) proposed to directly extract phrase pairs from comparable sentences. However, the number of retrievable phrase pairs is strongly limited, because one can collect such comparable sentences only on a relatively small scale for the targeted language pairs and domains.

When in-domain parallel or comparable sentences can not be easily retrieved, another possibility to find translations for OOV words is bilingual word lexicon induction using comparable or unaligned monolingual corpora (Fung, 1995; Rapp, 1995; Koehn and Knight, 2002; Haghighi et al., 2008; Daumé and Jagarlamudi, 2011; Irvine and Callison-Burch, 2013). This approach is especially useful in finding words and their translations specific to the given corpus. A recent and completely different trend of work uses an unsupervised method regarding translation as a decipherment problem to learn a bilingual word lexicon and use it as a translation model (Ravi and Knight, 2011; Dou and Knight, 2012; Nuhn et al., 2012). However, all these methods deal only with

[2] For instance, using these approaches on source and target monolingual data containing both 5 millions sentences means that we have to evaluate 25×10^{12} candidate sentence pairs.

words, mainly owing to the computational complexity of dealing with arbitrary lengths of phrases.

Translations of phrases can be induced using bilingual word lexicons and considering permutations of word ordering (Zhang and Zong, 2013; Irvine and Callison-Burch, 2014). However, it is costly to thoroughly investigate all combinations of a large number of word-level translation candidates and possible permutations of word ordering. To retain only appropriate phrase pairs, Irvine and Callison-Burch (2014) proposed to exploit a set of features. Some of them, including temporal, contextual, and topic similarity features, strongly relied on the comparability of Wikipedia articles and on the availability of news articles annotated with a timestamp (Klementiev et al., 2012). We may not have such useful resources in large quantity for the targeted language pairs and domains.

Saluja et al. (2014) and Zhao et al. (2015) also proposed methods to induce a phrase table, focusing only on the OOV words and phrases: unigrams and bigrams in the source side of their development and test data that are unseen in the training data. In their approach, no new translation options are proposed for known source phrases. To generate candidate phrase pairs, for a given source phrase, Saluja et al. (2014) uses only phrases from the target side of their parallel data and their morphological variants ranked and pruned according to the forward lexical translation probabilities given by their baseline system's translation model. Their approach thus strongly relies on the accuracy of the existing translation model. For instance, if the given source phrase contains only OOV tokens, as it may happen when translating a text from a different domain, their approach cannot retrieve candidate target phrases. Furthermore, they do not make use of external monolingual data to explore unseen target phrases. Their method is consequently inadequate to produce translations for phrases from a different domain than the one of the parallel data.

While Saluja et al. (2014) used a costly graph propagation strategy to score the candidate phrase pairs, Zhao et al. (2015) used a method with a much lower computational cost and reported higher BLEU scores using only word embeddings to score and rank many phrase pairs generated from target phrases, unigrams and bigrams, collected from monolingual corpora. The main contribution of Zhao et al. (2015) is the use of a local linear projection strategy (LLP) to obtain a cross-lingual semantic similarity score for each phrase pair. It makes the projection of source embeddings to the target embeddings space by learning a translation matrix for each source phrase embedding, trained on m gold phrase pairs with source phrase embeddings similar to the one to project. After the projection, based only on the similarity over embeddings, the k nearest target phrases of the projected source phrase are retrieved. If the projection for a given source phrase is not accurate enough, very noisy phrase pairs are generated. This may be a problem especially when the given source phrase does not need to be translated (i.e., numbers, dates, molecule names, etc.). The system will translate it, because this source phrase previously OOV is now registered in its induced phrase table, but has only wrong translations available (see Section 4.5 for empirical evidences).

3 In-domain phrase table induction

To induce an in-domain phrase table, our approach assumes the availability of large general-domain parallel data and in-domain monolingual data of both source and target languages. For some of our configurations, we also assume the availability of an in-domain bilingual lexicon to compute features associated with each candidate phrase pair and to compute a reliability score to filter appropriate ones.

3.1 In-domain phrase collection

In a standard configuration, SMT systems extract phrases of a length up to six or seven tokens. Collecting all the n-grams of such a length from a given large monolingual corpus is feasible, but will provide a large set of source and target phrases, resulting in an enormous number of candidate phrase pairs. In the next step, we evaluate each candidate in a given set of phrase pairs; it is thus crucial to get a reasonably small set of phrases.

In contrast with previous work, we collect more meaningful phrases than arbitrary short n-grams, using the following formula presented by Mikolov et al. (2013a):

$$score(w_i w_j) = \frac{freq(w_i w_j) - \delta}{freq(w_i) \times freq(w_j)}$$

where w_i and w_j are two consecutive tokens, $freq(\cdot)$ the frequency of a given word or phrase in the given monolingual corpus, and δ a discounting coefficient that prevents the retrieval of many phrases composed from infrequent words. Each bigram $w_i w_j$ in the monolingual corpus is scored with this formula and only the bigrams with a score above a predefined threshold θ are regarded as phrases. All the identified phrases are transformed into one token,[3] and a new pass is performed over the monolingual corpus to obtain new phrases also using the phrases identified in the previous passes. To further limit the number of collected phrases, we consider only phrases containing words that appear at least K times in the monolingual data. After T passes, we compile a set of phrases with (a) all the single words and (b) all the phrases with a length of up to L tokens identified during each pass.

Standard SMT systems for close languages directly output OOV tokens in the translation. To be as good as such systems, our approach must be able to retrieve the right translation, especially for the many domain-specific words and phrases that are identical in both source and target languages. To ensure that a source phrase that must remain untranslated has its identity in the target phrase set, we explicitly add in the target phrase set all the source phrases that also appear in the target monolingual data.

3.2 Feature engineering

Given two sets of phrases, for the source and target languages, respectively, we regard all possible combinations of source and target phrases as candidate phrase pairs. This naive coupling imperatively generates a large number of pairs that are mostly noise. Thus, the challenge here is to effectively estimate the reliability of each pair. This section describes several features to characterize each phrase pair; they are used for evaluating phrase pairs and also added in the induced phrase table to guide the decoder.

3.2.1 Cross-lingual semantic similarity

Many researchers tackled the problem of estimating cross-lingual semantic similarity between pairs of words or phrases by using their embeddings (Mikolov et al., 2013a; Chandar et al., 2014; Faruqui and Dyer, 2014; Coulmance et al., 2015; Gouws et al., 2015; Duong et al., 2016) in combination with either a seed bilingual lexicon or a set of parallel sentence pairs.

We estimate monolingual phrase embeddings *via* the element-wise addition of the word embeddings composing the phrase. This method performs well to estimate phrase embeddings (Mitchell and Lapata, 2010; Mikolov et al., 2013a), despite its simplicity and relatively low computational cost compared to state-of-the-art methods based on neural networks (Socher et al., 2013a; Socher et al., 2013b) or rich features (Lazaridou et al., 2015). This low computational cost is crucial in our case, as we need to evaluate a large number of candidate phrase pairs.

In order to make source and target phrase embeddings comparable, we perform a linear projection (Mikolov et al., 2013a) of the embeddings of source phrases to the target embedding space. To learn the projection, we use the method of Mikolov et al. (2013a) with the only exception that we deal with not only words but also phrases. Given training data, i.e., a gold bilingual lexicon, we obtain a translation matrix $\hat{W}$ by solving the following optimization problem with stochastic gradient descent:

$$\hat{W} = \arg\min_{W} \sum_{i} ||Wx_i - z_i||^2$$

where x_i is the source phrase embedding of the i-th training data, z_i the target phrase embedding of the corresponding gold translation, and W the translation matrix used to project x_i such that Wx_i is as close as possible to z_i in the target embedding space. One important parameter here is the number of dimensions of word/phrase embeddings. This can be different for the source and target embeddings, but must be smaller than the number of phrase pairs in the training data; otherwise the equation is not solvable. See Section 4.1 for the details about the bilingual lexicon used in our experiment.

Given a phrase pair to evaluate, the source phrase embedding is projected to the target embedding space, using $\hat{W}$. Then, we compute the cosine similarity between the projected source phrase embedding and the target phrase embedding to evaluate the semantic similarity between these phrases; this seems to give satisfying results in this cross-lingual scenario as shown by Mikolov et al. (2013a). A

[3] This transformation is performed by simply replacing the space between the two tokens with an underscore.

translation matrix is trained for each translation direction $f \rightarrow e$ and $e \rightarrow f$, respectively, so that we have two cross-lingual semantic similarity features for each phrase pair.

3.2.2 Lexical translation probabilities

We assume the existence of a large amount of general-domain parallel data, and train a regular translation model with lexical translation probabilities in an ordinary way. Although in-domain phrases are likely to contain tokens that are unseen in the general-domain parallel data, lexical translation probabilities may be useful to score candidate pair of source and target phrases that contain tokens seen in the general-domain parallel data. To compute a phrase-level score, for a target phrase e given a source phrase f, we consider all possible word alignments as follows:

$$P_{lex}(e|f) = \frac{1}{I}\sum_{i=1}^{I}\log\Big(\frac{1}{J}\sum_{j=1}^{J}p(e_i|f_j)\Big)$$

where I and J are the lengths of e and f, respectively, and $p(e_i|f_j)$ the lexical translation probability of the i-th target word e_i of e given the j-th source word f_j of f. Such phrase-level lexical translation probabilities are computed for both translation directions giving us two features.

3.2.3 Other features

As demonstrated by previous work (Irvine and Callison-Burch, 2014; Irvine and Callison-Burch, 2016), features based on the frequency of the phrases in the monolingual data may help us to better score a phrase pair. We add as features the inversed frequency of the source and target phrases in the in-domain monolingual data, along with their relative difference given by the following formula:

$$sim_f(e,f) = \left|\log\Big(\frac{freq(e)}{N_e}\Big) - \log\Big(\frac{freq(f)}{N_f}\Big)\right|$$

where N_x stands for the number of tokens in the in-domain monolingual data of the corresponding language.

The surface-level similarity of source and target phrases can also be a strong clue when considering the translation between two languages that are relatively close. We investigate two features concerning this: the first feature is the Levenshtein distance between the two phrases calculated regarding words as units,[4] while the other is a binary feature that fires if the two phrases are identical. We shall expect both features to be very useful in cases where many domain-specific words and phrases are written in the same way in two languages; for instance, drug and molecule names in the medical domain in French and English.

We also add as features the lengths of the source and target phrases, i.e., I and J, and their ratio.

Using all the above 12 features, the overall score for each pair is given by a classifier as described in Section 3.3; this score is also added as a feature in the induced phrase table for decoding.

3.3 Phrase pair filtering

As mentioned above, phrase pairs so far generated are mostly noise. To reduce the decoder's search space when using our induced phrase table, we radically filter out inappropriate pairs. Each candidate phrase pair is assessed by the method proposed in Irvine and Callison-Burch (2013), which predicts whether a pair of words are translations of one another using a classifier. As training examples, we use a bilingual lexicon as positive examples and randomly associated phrase pairs from our phrase sets as negative examples. For classification, we use all the features presented in Section 3.2.

We use the score given by the classifier to rank the target phrases for each source phrase. Only the target phrases with the top n scores are kept in the final induced phrase table.

4 Experiments

This section demonstrates the impact of the induced phrase tables in translating in-domain texts in three configurations. In the first configuration (Conf. 1), we evaluated whether our induced phrase table improves the translation of in-domain texts over the *vanilla* SMT system which used only one phrase table trained from general-domain parallel

[4]Here we did not use the character-level edit distance to measure the orthographic similarity between phrases. Even though such a feature may be useful (Koehn and Knight, 2002), its computational cost is too high to deal efficiently with billions of phrase pairs.

data. We then evaluated, in the second configuration (Conf. 2), whether our induced phrase table is also beneficial when used in an SMT system that already incorporates an in-domain bilingual lexicon that could be created manually or induced by some of the methods mentioned in Section 2. Finally, we evaluated in complementary experiments (Conf. 3) whether our induced phrase table can also offer useful information to improve translation quality even when used in combination with another standard phrase table generated from in-domain parallel data.

4.1 Data

Since our approach assumes the availability of large-scale general-domain parallel and monolingual corpora, we considered the French–English language pair and both translation directions for our experiments. The French–English version of the Europarl parallel corpus[5] was regarded as a general-domain, and not strictly out-of-domain, corpus because many debates can be associated to a specific domain and can contain phrases specific to particular domains. As general-domain monolingual data, we used the concatenation of one side of Europarl and the 2007–2014 editions of News Crawl corpora[6] in the same language.

We focused on two domains: medical (EMEA) and science (Science). For both domains, we used the development and test sets provided for a workshop on domain adaptation of MT (Carpuat et al., 2012).[7] We also used the provided in-domain parallel data for training but regarded only the target side as monolingual data. Since our primary objective is the induction of a phrase table without using in-domain parallel data, the source side of the in-domain parallel data was not used as a part of the source in-domain monolingual data, except when training an ordinary in-domain phrase table in Conf. 3. As medical domain monolingual data for the EMEA translation task, we used the French and English monolingual medical data provided for the WMT'14 medical translation task.[8] None of the parallel corpora provided for the WMT'14 medical translation task was used. As science domain monolingual data for the Science translation task, we used the English side of the ASPEC parallel corpus (Nakazawa et al., 2016).[9] Unfortunately, we did not find any French monolingual corpora publicly available for the Science domain that were sufficiently large enough for our experiments. Statistics on the data we used are presented in Table 1.

Domain	Data	# sent.	# tok. (En-Fr)
EMEA	development	2,022	28k-32k
	test	2,045	25k-29k
	parallel	472k	6M-7M
	monolingual		275M-255M
Science	development	1,990	52k-65k
	test	1,982	52k-65k
	parallel	66k	2M-2M
	monolingual		82M-2M
General	parallel	2M	54M-60M
	monolingual		2.8B-1.1B

Table 1: Statistics on train, development, and test data.

To induce the phrase tables from the monolingual data, we compared two bilingual lexicons: a general-domain and an in-domain lexicons. These lexicons are used to train the translation matrices (see Section 3.2.1) and to train the classifier (see Section 3.3). The general-domain lexicon (henceforth, `gen-lex`) is a phrase-based one extracted from the phrase table built on the general-domain parallel data (see Section 4.3). We extracted the 5,000 most frequent source phrases and their most probable translation according to the forward translation probability, $p(e|f)$. We adopted this size as it had been proven optimal to learn the mapping between two monolingual embedding spaces (Vulić and Korhonen, 2016). For some experiments, we also simulated the availability of an in-domain bilingual lexicon. We automatically generated a lexicon for each domain (henceforth, `in-lex`) using the entire in-domain parallel data, in the same manner as compiling `gen-lex`, except that we selected the 5,000 most frequent source words in the in-domain parallel data that were not in the 5,000 most frequent words in the general-domain parallel data in order

[5] `http://statmt.org/europarl/`, release 7

[6] `http://statmt.org/wmt15/translation-task.html`

[7] `http://hal3.name/damt/`

[8] `http://www.statmt.org/wmt14/medical-task/`

[9] `http://orchid.kuee.kyoto-u.ac.jp/ASPEC/`

Side	Domain	Data	w2p	w2v
source	general	monolingual		√
	in-domain	monolingual	√	√
		parallel	(√)	(√)
		development	√	
		test	√	
target	general	monolingual		√
	in-domain	monolingual	√	√
		parallel	√	√

Table 2: Corpora used for extracting phrases and computing word embeddings: w2p indicates word2phrase, while w2v for word2vec. (√) denotes that the data are used in Conf. 3 only.

Task		source all	source dev+test	target	# phrase pairs
EMEA	Fr→En	4.5M	20k	437k	8.7B
	En→Fr	5.1M	11k	469k	5.2B
Science	Fr→En	1.1M	28k	216k	6.0B
	En→Fr	2.3M	24k	18k	432M

Table 3: Size of the phrase sets collected from the source and target in-domain monolingual data and the number of phrases appearing only in the concatenation of the source side of the development and test sets (dev+test). "# phrase pairs" denotes the number of phrase pairs assessed by the classifier.

to ensure that we obtained mostly in-domain word pairs. Note that we did not use phrases but words for in-lex, assuming that humans are not able to manually construct a lexicon comprising phrase pairs similar to those in phrase tables for SMT systems. For Conf. 3, as we assume the availability of in-domain parallel data, the bilingual lexicon (para-lex) used was 5,000 phrase pairs extracted from the in-domain phrase table, excluding the source phrases of gen-lex.

4.2 Tools and parameters

A summary of the data used to collect phrases and estimate word embeddings is presented by Table 2.

For each pair of domain and translation direction, sets of source and target phrases were extracted from the in-domain monolingual data, as described in Section 3.1. As in previous work (Irvine and Callison-Burch, 2014; Saluja et al., 2014; Zhao et al., 2015), we focus on source phrases appearing in the development and test sets in order to maximize the coverage of our induced phrase table for them.[10] More precisely, source phrases were collected from the concatenation of the development and test sets and the in-domain monolingual data with reliable statistics, and then only phrases appearing in the development and test sets were filtered.[11] We removed phrases containing tokens unseen in the in-domain monolingual data, because we are unable to compute all our features for them. On the other hand, target phrases were collected from the in-domain monolingual data, including the target side of in-domain parallel data. To identify phrases, we used the word2phrase tool included in the word2vec package,[12] with the default values for δ and θ. We set $K = 1$ for the source language to ensure that most of the tokens would be translated, and $K = 25$ for the target language to limit the number of resulting phrases. We set $L = 6$ as this is the same maximal phrase length that we set for the phrase tables trained from the parallel data. We stopped at $T = 4$ passes as the fifth pass retrieved only a very small number of new phrases compared to the fourth pass. Statistics of the collected phrases for each task are presented in Table 3.

To train the word embeddings, we used word2vec with the following parameters: -cbow 1 -window 10 -negative 15 -sample 1e-4 -iter 15 -min-count 1. Mikolov et al. (2013a) observed that better results for cross-lingual semantic similarity were obtained when using word embeddings with higher dimensions

[10] We are aware that this may not be practical because it requires the knowledge of the development and test sets beforehand. For instance for the Fr→En EMEA translation task, inducing a phrase table given all the 4.5M collected source phrase would required approximately 3 months using 100 CPU threads. Increasing the value of K to collect less source phrases can be a reasonable alternative to significantly decrease this computation time, even though it will also necessarily decrease the coverage of the phrase table. We leave for our future work the study of a phrase table induction with source phrases extracted from source monolingual data without referring to the development and test sets.

[11] As we had no French monolingual corpus for the Science domain, the development and test sets for the Science Fr→En task were concatenated with one million sentences randomly extracted from the general-domain monolingual data.

[12] https://code.google.com/archive/p/word2vec/

Data	LM1	LM2	LM3
Target side of in-domain parallel data	√	√	√
In-domain monolingual data	√	√	
General-domain monolingual data	√		

Table 4: Source of our three language models.

Phrase table	Conf. 1	Conf. 2	Conf. 3
Phrase table trained from general-domain parallel data	√	√	√
Phrase table trained from in-domain parallel data			√
In-domain bilingual lexicon		√	
Phrase table induced from in-domain monolingual data	√	√	√

Table 5: Multiple phrase table configurations.

on the source side than on the target side. We therefore chose 800 and 300 dimensions for the source and target embeddings, respectively. The embeddings were trained on the concatenation of all the general-domain and in-domain monolingual data as presented by Table 2. Consequently, for each pair of domain and translation direction, we have four word embedding spaces: those with 300 or 800 dimensions for source and target languages.

The reliability of each phrase pair was estimated as described in Section 3.3 to compile phrase tables of reasonable size and quality. We used `Vowpal Wabbit`[13] to perform logistic regression with one pass, default parameters, and `--link logistic` option to obtain a classification score for each phrase pair. In the final induced phrase table, we kept the 300 best target phrases[14] for each source phrase according to this score.

4.3 SMT systems

The Moses toolkit (Koehn et al., 2007)[15] was used for training SMT models, parameter tuning, and decoding. The phrase tables were trained on the parallel corpus using SyMGIZA++ (Junczys-Dowmunt and Szał, 2012)[16] with IBM-2 word alignment and the `grow-diag-final-and` heuristics. To obtain strong baseline systems, all SMT systems used three language models[17] built on different sets of corpora as shown in Table 4; each language model is a 4-gram modified Kneser-Ney smoothed one trained using lmplz (Heafield et al., 2013).[18] To concentrate on the translation model, we did not use the lexical reordering model throughout the experiments, while we enabled distance-based reordering up to six words.

Our systems used the multiple decoding paths ability of Moses; we used up to three phrase tables in one system, as summarized in Table 5. We did not add the features presented in Section 3.2 to the phrase pairs directly derived from the parallel data.[19]

Weights of the features were optimized with `kb-mira` (Cherry and Foster, 2012) using 200-best hypotheses on 15 iterations. The translation outputs were evaluated with BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). The results were averaged over three tuning runs. The statistical significance was measured by approximate randomization (Clark et al., 2011) using `MultEval`.[20]

4.4 Additional baseline systems

To compare our work with a state-of-the-art phrase table induction method, we implemented the work of Zhao et al. (2015). Even though they did not propose their method to perform domain adaptation of an SMT system, their work is the closest to ours and does not require other external resources than those we used, i.e., parallel data and monolingual data not necessarily comparable. We implemented both global (GLP) and local (LLP) linear projection strategies and collected source and target phrases as they did. The source phrase set contains all uni-

[13] `https://github.com/JohnLangford/vowpal_wabbit/`

[14] As in Irvine and Callison-Burch (2014), we obtained better results when favoring recall over precision. We chose 300 empirically since we did not observe any further improvements when keeping more target phrases.

[15] `http://statmt.org/moses/`, version 2.1.1

[16] `https://github.com/emjotde/symgiza-pp/`

[17] The one exception is the system for the Science En→Fr task, which uses only two language models as we do not have any in-domain monolingual data in addition to the target side of the in-domain parallel data.

[18] `https://kheafield.com/code/kenlm/estimation`

[19] As in Irvine and Callison-Burch (2014), we got a drop of up to 0.5 BLEU points when we added our features, derived from monolingual data, to the original phrase table.

[20] `https://github.com/jhclark/multeval/`

grams and bigrams in the development and test sets, while the target phrase set contains unigrams and bigrams collected from the in-domain monolingual data. They did not mention any filtering of their phrase sets, but we chose to remove all phrases containing digits or punctuation marks, since trying to retrieve the translation of numbers or punctuation marks relying only on word embeddings seems inappropriate and in fact produced worse results in our preliminary experiments. To highlight the impact of the phrase sets used, we also experimented LLP using our phrase sets collected with `word2phrase`. Furthermore, to get the best possible results, we did not use the search approximations presented in Zhao et al. (2015), i.e., local sensitive hashing and redundant bit vector, and used instead linear search.

For the GLP configuration, the translation matrix was trained on `gen-lex`, i.e., 5,000 phrase pairs extracted from the general-domain phrase table trained on parallel data. For the LLP configurations, as in Zhao et al. (2015), we trained the translation matrix for each source phrase on the 500 most similar source phrases, retrieved from the general-domain phrase table, associated to their most probable translation. For both GLP and LLP configurations, we kept the 300 best target phrases for each source phrase. Four features, phrase and lexical translation probabilities for both translation directions, were approximated using the similarity between source and target phrase embeddings for each phrase pair and included in the induced phrase table as described by Zhao et al. (2015).

Since this approach proposes to translate all OOV unigrams and bigrams, it is likely in our scenario that some medical terms, for instance, will have no correct translations in the induced phrase table. For a comparison, we added one more baseline system, which merely uses a *vanilla* Moses with the `-du` option of Moses activated to drop all unknown words instead of copying them into the translation.

4.5 Results

The experimental results are given in Table 6.

In Conf. 1, our results show that both GLP and LLP configurations performed much worse than the *vanilla* Moses when using phrases naively collected. This is due to the fact that the induced phrase table contains translations for every OOV unigrams and bigrams, even for those who do not need to be translated, such as molecule names or place names. Word embeddings are well-known to be inaccurate for very infrequent words (Mikolov et al., 2013b); consequently, for some rare source phrases, even if the right translation is in the target phrase set, it is not guaranteed that it will be registered in the induced phrase table as one of the 300 best translations for the source phrase, relying only on word embeddings. The significant improvements over a *vanilla* Moses observed by Zhao et al. (2015) would potentially be because they translated from Arabic, and Urdu, to English. For such language pairs, one can safely try to translate every OOV token of a general-domain text, and it is unlikely to do worse than a *vanilla* Moses system that will leave the OOV tokens as is in the translation. As shown by the Moses `du` configurations, dropping them led to a drop of up to 4.2 BLEU points for the EMEA Fr→En translation task. This suggests that OOV tokens must be carefully translated only when necessary. Many OOV tokens in our translation tasks do not need to be translated into different forms. Hence, we regard the *vanilla* Moses that copies the OOV tokens in the translation a strong baseline system.

Interestingly, using the phrases collected by our method for LLP produced much better translations, even slightly better than the one produced by the *vanilla* Moses system for the EMEA En→Fr translation task with an improvement of 0.2 BLEU points. This may be due to the fact that our source phrase set is not only made from OOV phrases, meaning that new useful translations may be proposed for source phrases that are already registered in the general-domain phrase table. Moreover, with our phrase sets, the decoder also has the possibility to leave some tokens untranslated since we added each source phrase in the target phrase set if it appeared in the target monolingual data.

Instead of relying only on word embeddings, the features used in our approach helped significantly to improve the translation quality. When we added our induced phrase table to a *vanilla* Moses system, we observed consistent and significant improvements in translation quality, with up to 2.1 BLEU and 2.2 METEOR points of improvement for the Science En→Fr translation task.

Compared to the LLP method proposed by Zhao

Configuration	EMEA				Science			
	Fr→En		En→Fr		Fr→En		En→Fr	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
vanilla Moses `du`	24.2	27.4	21.7	40.0	22.3	29.1	20.4	42.7
vanilla Moses (Conf. 1)	28.4	30.1	25.4	44.8	24.1	30.4	22.7	45.1
+ GLP IPT naive	24.3	27.0	22.3	41.0	22.4	29.0	20.6	42.6
+ LLP IPT naive	24.7	27.4	22.0	40.4	22.5	29.3	21.1	43.4
+ LLP IPT	27.9	29.6	25.6	**45.1**	22.7	29.3	21.3	43.5
+ our IPT (`gen-lex`)	**30.2**	**32.1**	**27.1**	**46.6**	**25.4**	**32.0**	**24.8**	**47.3**
+ in-domain bilingual lexicon (Conf. 2)	32.4	32.8	28.3	48.2	26.6	32.4	24.9	48.0
+ our IPT (`gen-lex`)	**33.5**	32.6	**28.8**	**48.6**	**28.5**	**33.8**	**25.2**	**48.4**
+ our IPT (`in-lex`)	**33.8**	32.9	**29.2**	**48.9**	**26.9**	**32.7**	**25.9**	**49.0**
+ in-domain phrase table (Conf. 3)	39.1	36.1	33.8	53.1	32.1	36.1	31.0	53.9
+ our IPT (`para-lex`)	39.1	36.1	34.0	53.2	32.1	36.1	31.2	54.1

Table 6: Results (BLEU and METEOR) with an induced phrase table (IPT). The Moses `du` and *vanilla* Moses systems use only one phrase table trained from the general-domain parallel data. The translation matrices and the classifiers have been trained with a bilingual lexicon: `gen-lex`, `in-lex`, or `para-lex`. The configurations denoted as "naive" use a phrase table induced from phrases collected as described in Section 4.4. Bold scores indicate the statistical significance ($p < 0.01$) of the gain over the baseline system (Conf. X) in each configuration.

et al. (2015), our approach includes more features and an additional classification step. Thus, the induction of a phrase table is much slower. For instance, for the EMEA Fr→En translation task, using the phrase sets extracted with `word2phrase`, our induction method (excluding phrase collection) was nearly 14 times slower (9 hours vs. 38 minutes).[21] Phrase collection using `word2phrase` was much faster than feature computation and phrase pair classification. For instance, it took 72 minutes to collect target phrases for the EMEA Fr→En translation task, using four iterations of `word2phrase` on the English in-domain monolingual data with 1 CPU thread.

In Conf. 2, adding an in-domain bilingual lexicon as a phrase table to the *vanilla* baseline system significantly boosted the performance, mainly by reducing the number of OOV tokens. Our induced phrase tables had less impact, probably due to the overlap between useful word pairs contained in both the induced phrase table and the added bilingual lexicon. However, we still observed significant improvements, which support the usefulness of the induced phrase table, with up to 1.4 and 1.0 BLEU points of improvements, respectively, for the EMEA Fr→En and Science En→Fr translation tasks for instance. In this configuration, the `in-lex` phrase table led to slight but consistent improvements. It helped more than the `gen-lex` phrase table, except in the Science Fr→En task, for which the use of the `gen-lex` phrase table yielded significantly better results than the use of the `in-lex` phrase table. We can expect such differences when the classifier and the translation matrices are trained using infrequent words. Embeddings for such words are typically not as well estimated as those for frequent words, meaning that the features based on the word embeddings are less reliable and thus mislead both the classifier for pruning and the decoder.

In Conf. 3, where the baseline system even used a phrase table trained on in-domain parallel data, we obtained contrasted results, with only slight improvements for the En→Fr translation direction and no improvements for the Fr→En translation direction. This lack of improvement may be due to the more reliable features and more accurate phrase pairs contained in the phrase table directly learned from the parallel data. This may lead the decoder to prefer this table to the induced one and give higher weights of its features according to this preference during tuning.

[21] The experiments were performed with 20 CPU threads. Note also that computational speed was not our primary focus when implementing our approach. Optimizing our implementation may lead to significant gains in speed, while Zhao et al. (2015) have presented a search approximation able to make their approach 18 times faster than linear search.

	EMEA				Science			
	Fr→En		En→Fr		Fr→En		En→Fr	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
correct	53.1	55.1	52.8	54.2	54.0	57.1	55.3	57.2
SEEN	6.6	3.9	7.8	6.1	5.9	2.3	8.5	2.2
SENSE	18.1	13.3	15.6	11.5	18.9	14.0	12.9	12.9
SCORE	22.2	27.7	23.8	28.2	21.2	26.6	23.3	27.7

Table 7: Percentage of the source tokens: comparison of the translations generated with (w/) or without (w/o) our gen-lex induced phrase table (Conf. 1).

5 Error analysis

In Section 5.1, we first present an analysis of the distribution of translation errors that our systems produced, using the S^4 taxonomy (Irvine et al., 2013). Then, in Section 5.2, we illustrate some translation examples for which our induced phrase tables have produced a better translation.

5.1 Analysis with the S^4 taxonomy

The S^4 taxonomy comprises the following four error types:

- SEEN: attempt to translate a word never seen before
- SENSE: attempt to translate a word with the wrong sense
- SCORE: a good translation for the word is available but another one, giving a better score to the hypothesis, is chosen by the system
- SEARCH: a good translation is available for the word but is pruned during the search for the best hypothesis

We considered the SEEN, SENSE, and SCORE errors as in Irvine et al. (2013), but not the SEARCH errors, assuming that the recent phrase-based SMT systems rarely make this type of errors and without impact on the translation quality (Wisniewski et al., 2010; Aziz et al., 2014). We performed a Word Alignment Driven Evaluation (WADE) (Irvine et al., 2013) to count the word-level errors.

Table 7 compares the results with and without our gen-lex induced phrase tables (Conf. 1). For the four tasks, more than half of the source tokens were correctly translated according to the translation reference. Our analysis reveals that our induced phrase table helps to obtain more correct translations, as higher percentages of source words were correctly translated, despite the significant increase of SCORE errors (around 5% for all the tasks). This means that the correct translation for the source word is available, but the features associated to this translation were not informative enough for the decoder to choose it. The percentage of SEEN errors in the translations decreased significantly with the induced phrase table for all the tasks, as a result of many words and phrases unseen in the general-domain parallel data being covered by using the in-domain monolingual data. However, our method does not guarantee to find appropriate translations for these words. It is even possible that all the proposed translations are inappropriate. Nonetheless, we can see a noticeable decrease of the SENSE errors, except in the Science En→Fr task, for which we have used only a small amount of in-domain French monolingual data. As reported in Table 3, fewer target phrases were collected for this task, leading to only a small chance of obtaining the right translation for a given source phrase. The percentage of SENSE errors still remains higher than 10% for all tasks, indicating that the correct translation is not available in our phrase set or is pruned by the classifier during the phrase table induction.

From this analysis, we draw the conclusion that our approach has significantly increased the reachability of the translation reference along with the quality of the translation produced by the decoder. We expect that more informative or better estimated features can further improve our results. Improving our method to collect the target phrases or using a larger in-domain monolingual corpus would also help to reduce SENSE errors.

5.2 Translation examples

Table 8 presents examples of source phrase and their translations chosen by the decoder in the EMEA Fr→En translation task. As shown by Example #1, both LLP and gen-lex configurations can find a good translation in their induced phrase table for the phrase "au point d'injection" while the general-domain phrase table does not contain this source phrase. As a result, the *vanilla* Moses system produced a wrong translation using general-domain word translations.

System	#1	#2	#3	#4
source	au point d' injection	glaucome aigu	contient du lactose monohydraté	le lansoprazole n' est pas
vanilla Moses	at injection	acute glaucome	monohydraté contains lactose	the lansoprazole is not
LLP IPT	at the point of injection	acute	contains lactose	the , is not
our IPT (`gen-lex`)	at the site of injection	acute glaucoma	contains lactose monohydrate	the lansoprazole is not
reference	at the injection site	acute glaucoma	contains lactose monohydrate	lansoprazole is not

Table 8: Examples of source phrase and their translation, from the test set of the EMEA Fr→En translation task, produced by the decoder using different configurations: *vanilla* Moses (Conf. 1) and Moses using a phrase table induced with LLP or with our method (`gen-lex`).

Example #2 shows a typical error made by the LLP configuration. In this example, "glaucome" is OOV, no translation is proposed for this token in the general-domain phrase table. The LLP IPT contains the source phrase "glaucome aigu" but none of the 300 best corresponding target phrases contain the token "glaucoma". However, most of them contain the meaning of "acute". This can be explained by the much higher frequency of "aigu" while the word "glaucome" is very rare, even in the in-domain monolingual data. Consequently, "aigu" has an embedding more accurate than the one of "glaucome" which is then much more difficult to project correctly across languages. In contrast, our `gen-lex` IPT contains the translation reference for "glaucome aigu" and this translation has been used correctly by the decoder, guided by our feature set.

Example #3 is similar to Example #2, the embedding of the rare word "monohydraté" is probably not accurate enough to be correctly projected, the correct translation is not in the LLP IPT, while our approach succeeded to translate it correctly.

Finally, Example #4 presents another common situation where an OOV token, here "lansoprazole" has to be preserved as is and is correctly reported in the translation by the *vanilla* Moses system. The LLP IPT proposes translations for "lansoprazole", most of them semantically unrelated, like the one chosen by the decoder in this configuration.

We assume that the surface-level similarity features of our method helped the decoder to identify the right translation in this situation. Nonetheless, even when using our `gen-lex` IPT, we still observed some situations where tokens that should be preserved were actually wrongly translated, producing outputs worse than those produced by the *vanilla* Moses system.

6 Conclusion and future work

We presented a framework to induce a phrase table from unaligned monolingual data of specific domains. We showed that such a phrase table, when integrated to the decoder, consistently and significantly improved the translation quality for texts in the targeted domain. Our approach uses only simple features without requiring strongly comparable or annotated texts in the targeted domain.

Our method could further be improved in several ways. First, we expect better improvements by using more in-domain monolingual data or by being more careful in collecting the target phrases to use for the phrase table induction as opposed to simply pruning them according to the word frequency. Moreover, as we saw in Section 5, scoring the phrase pairs is one of the most important issues. We need more informative features to better score the pairs of source and target phrases. Despite their high computational cost, including features based on orthographic similarity or using better estimated cross-lingual embeddings may help for this purpose.

Acknowledgments

We would like to thank the anonymous reviewers and the action editor, Chris Quirk, for their insightful comments.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of EMNLP*, Edinburgh, Scotland, UK.

Wilker Aziz, Marc Dymetman, and Lucia Specia. 2014. Exact Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of EMNLP*, Doha, Qatar.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, et al. 2012. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins summer workshop final report. Baltimore, MD: Johns Hopkins University*.

A. P. Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2014. An Autoencoder Approach to Learning Bilingual Word Representations. In *Proceedings of NIPS*, Montréal, Canada.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of NAACL-HLT*, Montréal, Canada.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of ACL-HLT*, Portland, OR, USA.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, Fast Cross-lingual Word-embeddings. In *Proceedings of EMNLP*, Lisbon, Portugal.

Hal Daumé, III and Jagadeesh Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words. In *Proceedings of ACL-HLT*, Portland, OR, USA.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of EACL*, Gothenburg, Sweden.

Qing Dou and Kevin Knight. 2012. Large Scale Decipherment for Out-of-domain Machine Translation. In *Proceedings of EMNLP-CoNLL*, Jeju Island, Korea.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *Proceedings of EMNLP*, Austin, TX, USA.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of EACL*, Gothenburg, Sweden.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proceedings of EMNLP*, Barcelona, Spain.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*, Cambridge, MA, USA.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of ICML*, Lille, France.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *Proceedings of ACL-HLT*, Colombus, OH, USA.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL*, Sofia, Bulgaria.

Sanjika Hewavitharana and Stephan Vogel. 2016. Extracting parallel phrases from comparable data for machine translation. *Natural Language Engineering*, 22(4):549–573.

Ann Irvine and Chris Callison-Burch. 2013. Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals. In *Proceedings of HLT-NAACL*, Atlanta, GA, USA.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating Phrase Translations for Low Resource MT. In *Proceedings of CoNLL*, Baltimore, MD, USA.

Ann Irvine and Chris Callison-Burch. 2016. End-to-End Statistical Machine Translation with Zero or Small Parallel Texts. *Natural Language Engineering*, 22(4):517–548.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring Machine Translation Errors in New Domains. *Transactions of the Association for Computational Linguistics*, 1.

Marcin Junczys-Dowmunt and Arkadiusz Szał. 2012. SyMGiza++: Symmetrized Word Alignment Models for Machine Translation. In *Security and Intelligent Information Systems (SIIS)*, volume 7053 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin/Heidelberg, Germany.

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward Statistical Machine Translation without Parallel Corpora. In *Proceedings of EACL*, Avignon, France.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, Philadelphia, PA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, Prague, Czech Republic.

Angeliki Lazaridou, Georgiana Dinu, Adam Liska, and Marco Baroni. 2015. From Visual Attributes to Adjectives through Decompositional Distributional Semantics. *Transactions of the Association for Computational Linguistics*, 3.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, Lake Tahoe, NV, USA.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8).

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proceedings of LREC*, Portorož, Slovenia.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering Foreign Language by Combining Language Models and Context Vectors. In *Proceedings of ACL*, Jeju Island, Korea.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, PA, USA.

Reinhard Rapp. 1995. Identifying Word Translations in Non-parallel Texts. In *Proceedings of ACL*, Cambridge, MA, USA.

Sujith Ravi and Kevin Knight. 2011. Deciphering Foreign Language. In *Proceedings of ACL-HLT*, Portland, OR, USA.

Avneesh Saluja, Hany Hassan, Kristina Toutanova, and Chris Quirk. 2014. Graph-based Semi-Supervised Learning of Translation Models from Monolingual Data. In *Proceedings of ACL*, Baltimore, MD, USA.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013a. Parsing with Compositional Vector Grammars. In *Proceedings of ACL*, Sofia, Bulgaria.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of EMNLP*, Seattle, WA, USA.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of ACL*, Sapporo, Japan.

Ivan Vulić and Anna Korhonen. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of ACL*, Berlin, Germany.

Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. 2010. Assessing Phrase-Based Translation Models with Oracle Decoding. In *Proceedings of EMNLP*, Cambridge, MA, USA.

Jiajun Zhang and Chengqing Zong. 2013. Learning a Phrase-based Translation Model from Monolingual Data with Application to Domain Adaptation. In *Proceedings of ACL*, Sofia, Bulgaria.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of IEEE ICDM*, Maebashi, Japan.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning Translation Models from Monolingual Continuous Representations. In *Proceedings of NAACL-HLT*, Denver, CO, USA.